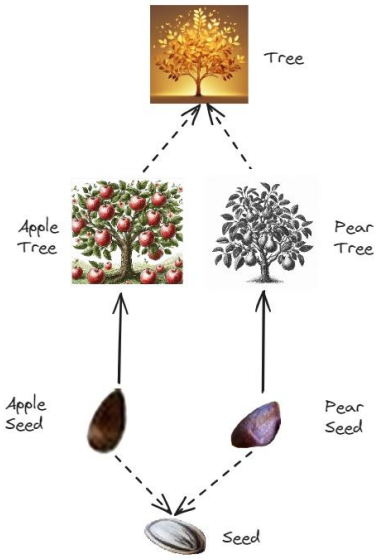
Reinventing analytic systems to unify and analyze the world's data - Part 2

Partha Nageswaran, Founder & CEO, Aabhra Inc (partha@aabhra.com)

A 4-year-old can accurately predict that an apple seed planted in an orchard of pear trees will still grow into an apple tree[(1)].



Even at this young age, the child has built a sound understanding of **ontological notions** such as

- Seeds, trees, fruits
- Their different types and relationships across each other

into their cognitive skills. Leveraging these, the child analyzes and predicts the outcome.

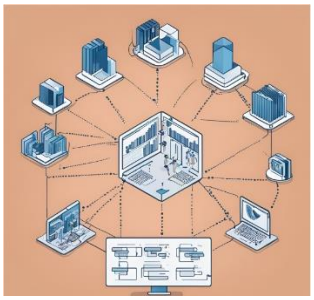
Ontologies provide structural, semantic, relational and behavioral definitions of things.

- The **Structure** of a tree such as: it has roots, trunks, branches, leaves.
- Its **Semantics** such as: it is anchored by roots; it generally grows over a few feet high.
- Its **Relationships** to other things, such as: to Seeds, to Fruits, etc.
- Its **Behavior**, such as: how its flowers may be pollinated, self or cross, etc.

Data, too, has structural, semantic, relational and behavioral characteristics grounded in ontologies. These are genetic characteristics of the data.

Data, data everywhere:

Enterprises and analysts are getting inundated by complex, voluminous data[2][3].



We need to build systems that leverage ontologies, much like humans do, to automatically understand data.

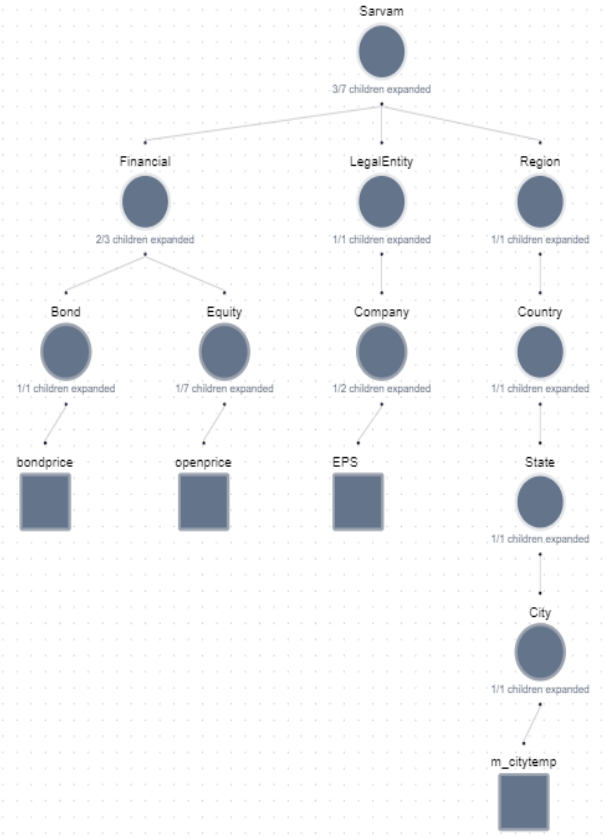
Such systems, using AI, can then help analyze data at scale, in a timely and accurate manner.

Ontologies and Data:

Using AI, genetic properties of data can be inferred quickly and used to naturally classify the data in an ontology.

Without lifting or shifting data, an ontological view of the data can be created automatically and rapidly.

Sarvam in Sanskrit means "everything".



Ontologies and Data Analysis - Structural:

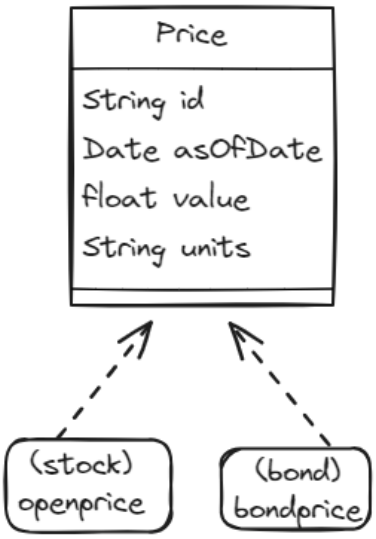
What does it mean to us when something is quoted as a price of an item? For example, what does it mean to us that the opening price (openprice) of a stock is $101.73?

Looking at the **Notion** of **Price** from an ontological perspective, structurally it has 4 characteristics at a minimum:

- Identifier of the thing that is priced.
- Numeric value of the price.
- Observation or as of date.
- Currency/units.

This is why we select the id, observation date and currency/units along with the value when we write SQL to retrieve price data.

Without any one of them, it is not proper pricing data.

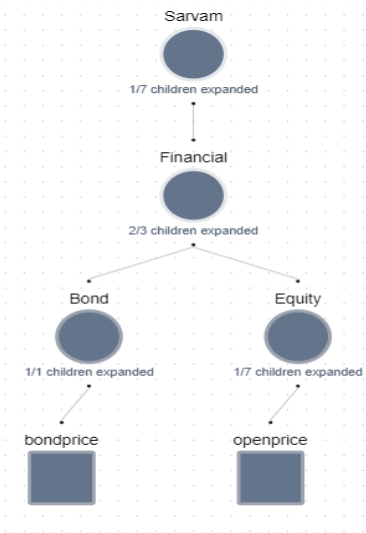


What are the genetic characteristics of an average of price over 4 days?

- For example, what is the average of the observation dates of the 4 days for which price points are averaged? Or does the average price not have an observation date?
- How does this impact interactions between the derived data and other data, including time-series data?

Today analysts define all of these over and over again in code, SQL, metadata and via other digital janitorial means, during each analysis.

Ontologies and Data Analysis - Semantical:



Consider the data in the context of an ontology.

Can you add bondprice to a stock's openprice?

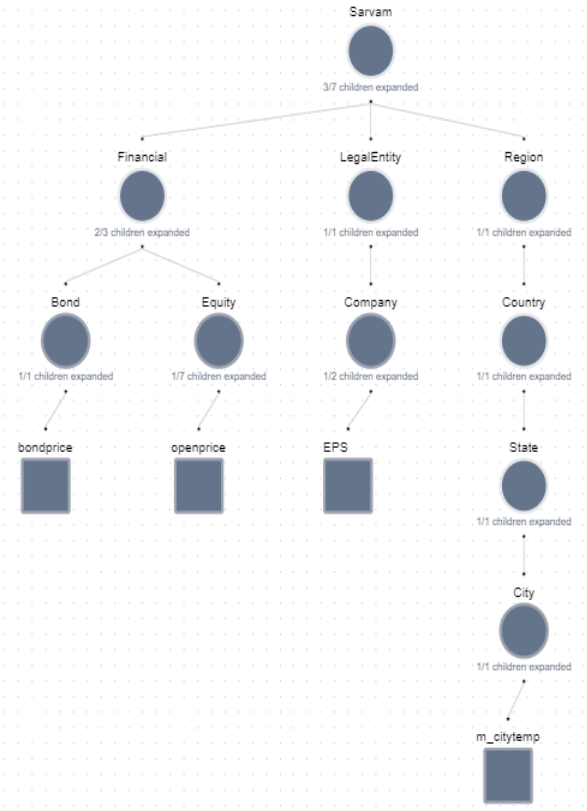
- Mathematically, yes (if the units are the same), but is the result meaningful?
- If so, what are the genetic characteristics of the result

Analytic systems must automatically:

- Understand data and its genetic characteristics.
- Derive not just the resultant values, but also genetic characteristics of the result.

Analysis made easy, basic examples:

How can we analyze, say, the Pearson correlation of the bondprice and (stock) openprice, by expressing just the intent?



Example 1: Pair-wise correlation of price of a stock or list of stocks and prices of all bonds *issued by the same issuer of the stock(s),* over the last 200 days.

This is correlation from the **ontological perspective** of Equity (see Sarvam ontology).

eval(PearsonR(openprice(-199d), bondprice(-199d)), ['stock1', 'stock2'])

Note: The list of stocks could be replaced by a complex screening expression.

That's it!

This should return the pair-wise Pearson R for <stock1, bond1>, <stock1, bond2>, …, <stock2, bond1>, <stock2, bond2>, …

Note that given a list of stocks the corresponding bonds issued by the issuer of the stocks should be automatically inferred from predefined ontological relations & AI.

This approach facilitates enmasse factor or signal identification, easily.

Pearson Correlation

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\,n\Sigma x^2 - (\Sigma x)^2\,]\,[\,n\Sigma y^2 - (\Sigma y)^2\,]}}$$

```
{
  nSigmaXY(x, y) = mult(count($x), add(mult($x, $y)));

  prodSigmaXSigmaY(x, y) = mult(add($x), add($y));

  numer(x, y) = diff(nSigmaXY($x, $y), prodSigmaXSigmaY($x, $y));

  denom(x, y) = sqrt(mult(numer($x, $x), numer($y, $y)));

  def PearsonR(x, y) = div(numer($x, $y), denom($x, $y));
}
```

To define PearsonR or any other formula, one should **simply transcribe the formula**.

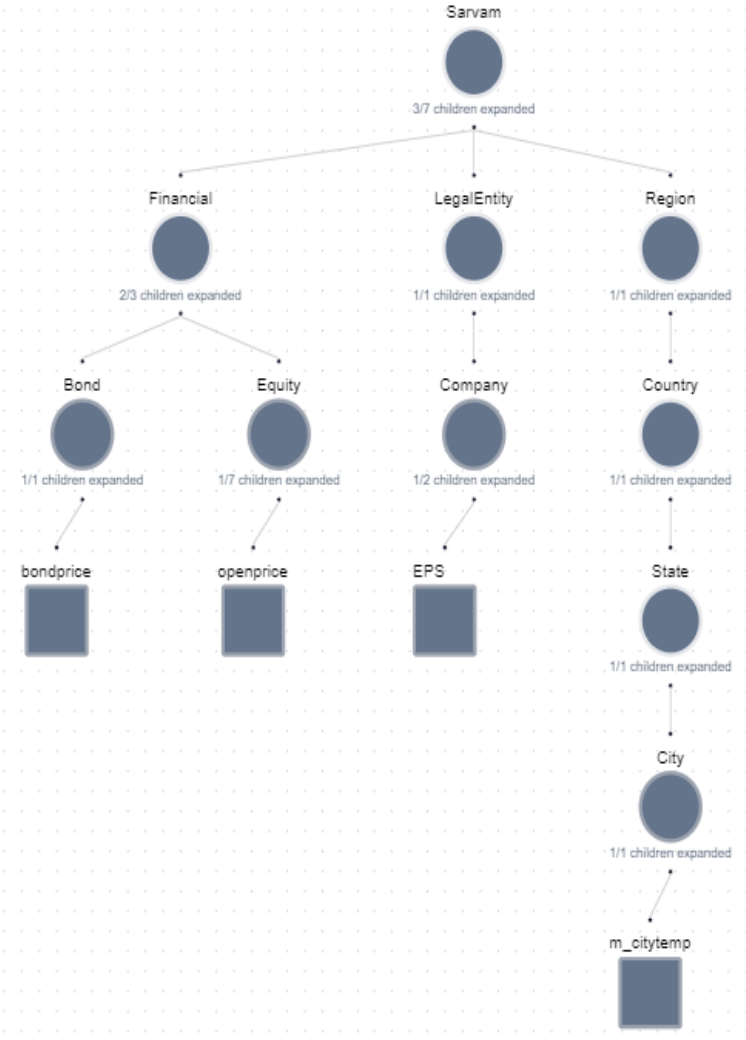
And easily compose these into very complex models using standard operators.

Example 2: Pair-wise correlation of prices of all bonds and those of all stocks issued by issuer(s).

This is similar to example 1 above, with an expanded **ontological perspective** (Company perspective, see Sarvam ontology).

eval(PearsonR(openprice(-199d), bondprice(-199d)), ['corp1', 'corp2'])

which would return, for each company, all the pair-wise correlations between the prices of all bonds and all stock issued by that company.



Example 3: Further expanding the scope of analysis, the ontological perspective could be City (of domicile, for example, see Sarvam ontology).

eval(PearsonR(openprice(-199d), bondprice(-199d)), ['city1'])

which would give pair-wise correlations between prices of all stocks and bonds of each company, for all companies domiciled in that city.

Avoid Data Preparation:

Data cleansing & preparation should be dynamic, enabling easy what-if's without creating multiple versions of data on disk.

E.g.: replaceNullsWith0s(openprice(-199d)), winsorize(openprice(-199d), …), etc. or compositions of pre-defined or user-defined functions used dynamically during analysis.

Summary:

Analysts should simply be able to focus on analysis by:

- Easily transcribing & using analytic functions or models of any complexity.
- Defining scope by choosing ontological perspective(s) for their analysis.
- Referencing any data for the analysis, from any domain, in the context of an ontology.
- Relying on the platform to flag semantically inaccurate operations.

That future is already here:

To find out more about unifying the world's data and performing intent driven analysis using AI & ontologies, without having to code or perform digital janitorial work, drop us a note at enquire@aabhra.com.

(1) https://helendecruz.net/docs/DeCruz_DeSmedt_BiolPhil.pdf
(2) https://ieeexplore.ieee.org/abstract/document/9862807
(3) https://www.anaconda.com/resources/whitepapers/state-of-data-science-report-2022/